The Efficacy of Human Post-Editing for Language Translation

Spence Green, Jeffrey Heer, and **Christopher D. Manning**
Computer Science Department, Stanford University
{spenceg,jheer,manning}@stanford.edu

ABSTRACT

Language translation is slow and expensive, so various forms of machine assistance have been devised. Automatic machine translation systems process text quickly and cheaply, but with quality far below that of skilled human translators. To bridge this quality gap, the translation industry has investigated *post-editing*, or the manual correction of machine output. We present the first rigorous, controlled analysis of post-editing and find that post-editing leads to reduced time and, surprisingly, improved quality for three diverse language pairs (English to Arabic, French, and German). Our statistical models and visualizations of experimental data indicate that some simple predictors (like source text part of speech counts) predict translation time, and that post-editing results in very different interaction patterns. From these results we distill implications for the design of new language translation interfaces.

Author Keywords

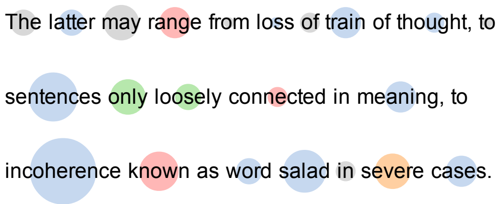
Language translation, post-editing, experiment, modeling

ACM Classification Keywords

H.5.2 Information Interfaces: User Interfaces; I.2.7 Natural Language Processing: Machine Translation

INTRODUCTION

High quality language translation is expensive. For example, the entire CHI proceedings from 1982 to 2011 contain 2,930 papers. Assuming roughly 5,000 words per paper and \$0.15 per word—a representative translation rate for technical documents—the cost to translate the proceedings from English to just one other language is \$2.2 million. Imagine: this sum is for one conference in one subfield of one academic discipline. To lower this cost, various forms of machine assistance have been devised: *source* (input) aids like bilingual dictionaries; *target* (output) aids such as spelling and grammar checkers; and *post-editing* (see [2]), the manual correction of fully automatic machine translation (MT) output. Language translation in practice is thus fundamentally an HCI task, with humans and machines working in concert.



(a) English input sentence with mouse hover visualization

MT:	Celui-ci peut aller de la perte d'un train de la pensée
Post-edit:	**Ceux-ci peuvent** aller de la perte **du fil** de la pensée

(b) Post-editing of French MT output

Figure 1: Translation as post-editing. (a) Mouse hover events over the source sentence. The color and area of the circles indicate part of speech and mouse hover frequency, respectively, during translation to French. Nouns (blue) seem to be significant. (b) The user corrects two spans in the MT output to produce a final translation.

Fully automatic MT is almost free, but the output, as represented by state-of-the-art systems such as Google Translate and Bing Translator, is useful for *gisting*—obtaining a rough idea of the translation—but far below the quality of skilled human translators. Nevertheless, the ostensible cost and speed benefits of MT are too appealing to resist, so the translation industry has long incorporated post-editing functionality into translator interfaces. But in terms of translation time and quality—the two variables of primary interest—post-editing has a mixed track record both quantitatively [46, 20] and qualitatively [32, 48]. Some studies have shown decreased translation time but lower quality, and even if speed does increase, translators often express an intense dislike for working with MT output.

This paper presents a controlled experiment comparing post-editing (hereafter "post-edit") to unaided human translation (hereafter "unaided") for three language pairs. We test four hypotheses: (1) post-edit reduces translation time, (2) post-edit increases quality, (3) suggested translations prime the translator, and (4) post-edit results in less drafting (as measured by keyboard activity and pause duration). Our results clarify the value of post-editing: it decreases time and, surprisingly, improves quality for each language pair. It also seems to be a more passive activity, with pauses (as measured by input device activity) accounting for a higher proportion of the total translation time. We find that MT suggestions prime translators but still lead to novel translations, suggesting new possibilities for re-training MT systems with human corrections.

CHI 2013, April 27–May 2, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-1899-0/13/04...\$15.00.

Our analysis suggests new approaches to the design of translation interfaces. "Translation workbenches" like the popular SDL Trados package are implicitly tuned for translation drafting, with auto-complete dialogs that appear while typing. However, when a suggested translation is present, we show that translators draft less. This behavior suggests UI designers should not neglect modes for source comprehension and target revision. Also, our visualizations (e.g., Figure 1) and statistical analysis suggest that assistance should be optimized for certain parts of speech that affect translation time.

We first review prior work on post-editing and machine-assisted translation. Then we describe the translation experiment. Next, we present visual and statistical analyses of the new translation data. Finally, we correlate the visual and statistical results with user feedback, and distill design implications.

RELATED WORK

Translation is a difficult computational task because it is hard to routinize. Consequently, the idea of combining human and machine expertise (see [8, 28]) was one avenue for re-starting machine translation research, which had stalled in the mid-1960s. Industry saw post-editing as one way to aid translators with imperfect contemporary technology, with conferences on the subject convened at least as early as 1981 [33]. Unfortunately, the HCI perspective on MT has been overlooked in natural language processing (NLP), where end-to-end automation of the translation process has been the preeminent goal since the statistical revolution in the early 1990s [35].

This paper unites three threads of prior work: visual analysis of the translation process, bilingual post-editing, and monolingual collaborative translation.[1]

Visual Analysis of the Translation Process

Post-editing involves cognitive balancing of source text comprehension, suggested translation evaluation, and target text generation. When interface elements are associated with these processes, eye trackers can give insight into the translation process. O'Brien [41] used an eye tracker to record pupil dilation for post-editing for four different source text conditions, which corresponded to percentage match with a machine suggestion. She found that pupil dilation, which was assumed to correlate with cognitive load, was highest for the no assistance condition, and lower when any translation suggested was provided.

Carl and Jakobsen [14] and Carl et al. [15] recorded fixations and keystroke/mouse activity. They found the presence of distinct translation phases, which they called *gisting* (the processing of source text and formulation of a translation sketch), *drafting* (entry of the translation), and *revision*, in which the draft is refined. Fixations clustered around source text during gisting, the target text entry box during drafting, and in both areas during revision.

In practice, eye trackers limit the subject sample size due to convenience and cost. We will track mouse cursor movements as a proxy for focus. This data is easy to collect, and correlates with eye tracking for other UIs [16, 26], although we do not explicitly measure that correlation for our task.

Bilingual Post-editing

The translation and NLP communities have focused largely on bilingual post-editing, i.e., the users are proficient in both source and target languages. Krings [31] conducted early work[2] on the subject using the Think Aloud Protocol (TAP), in which subjects verbalize their thought processes as they post-edit MT output. He found that the post-edit condition resulted in a 7% decrease in time over the unaided condition on a paper medium, but a 20% increase in time on a computer screen. However, Krings [31] also observed that TAP slowed down subjects by nearly a third.

Later work favored passive logging of translator activity. O'Brien [39] used Translog [27], which logs keyboard and mouse events, to measure the effect of source features on time. Subsequently, O'Brien [42] investigated the hypothesis that longer duration pauses reflect a higher cognitive burden (see [45]) and thus slower translation time. However, both of these experiments focused on the effect of rule-based, language-specific source features (see [7]). For instance, "words that are both adverbs and subordinate conjunctions" (e.g., 'before') were selected. The generality of such rules is unclear.

Guerberof [22] focused instead on comparison of post-edit to unaided. She observed reduced translation time with post-editing—albeit with very high per subject variance—but slightly lower quality according to a manual error taxonomy. However, she used only nine subjects, and did not cross source sentences and conditions, so it was not possible to separate sentence-specific effects.

In contrast, Koehn [29] crossed five blocks of English-French documents with five different translation conditions: unaided, post-edit, and three different modes of interactive assistance. He used 10 student subjects, who could complete the experiment at any pace over a two week period, and could use any type of alternate machine assistance. He found that, on average, all translators produced better and faster translations for the four assisted conditions, but that the interactive modes offered no advantage over post-editing.

Results derived from small data samples and student subjects may not generalize to industrial settings. At Adobe, Flournoy and Duran [19] found that post-editing resulted in a 22-51% decrease in translation time for a small scale task (about 2k source tokens) and 40-45% decrease for a large-scale task (200k source tokens). They also found that MT quality varied significantly across source sentences, with some translations requiring no editing and others requiring full re-translation. Likewise, at Autodesk, Plitt and Masselot [44] found that post-editing resulted in a 74% average reduction in time. Quality was assessed by their corporate translation staff using an unpublished error classification method. The raters found a lower error rate in the post-edit condition.

These large-scale experiments suggested that post-editing reduces time and increases quality. However, at Tilde, Skadiņš et al. [46] also observed reduced translation time for post-edit,

[1] See Tatsumi [47] for a broader survey of post-editing.

[2] Krings [31] is an English translation of the 1994 thesis, which is based on experiments from 1989-90.

but with a higher error rate for all translators. Like the other industrial studies, they did not report statistical significance.

Garcia [20] was the first to use statistical hypothesis testing to quantify post-editing results. In the larger of his three experiments, he measured time and quality for Chinese-English translation in the unaided vs. post-edit conditions. Statistically significant improvements for both dependent variables were found. Smaller experiments for English-Chinese translation using an identical experimental design did not find significant effects for time or quality. These results motivate consideration of sample sizes and cross-linguistic effects.

Finally, Tatsumi [47] made the only attempt at statistical prediction of time given independent factors like source length. However, she did not compare her regression model to the unaided condition. Moreover, her models included per-subject factors, thus treating subjects as fixed effects. This choice increases the risk of type II errors when generalizing to other human subject samples.

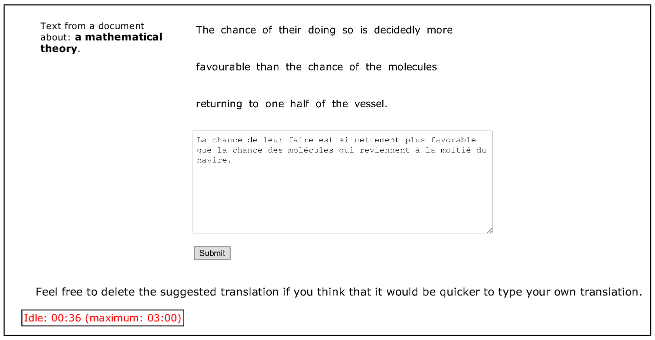
Monolingual Collaborative Translation

In contrast to bilingual post-editing, the HCI community has focused on *collaborative translation*, in which monolingual speakers post-edit human or machine output.[3] Quality has been the focus, in contrast to bilingual post-editing research, which has concentrated on time. Improvements in quality have been shown relative to MT, but not to translations generated or post-edited by bilinguals.

Morita and Ishida [37, 38] proposed a method for partitioning a translation job between source speakers, who focus on adequacy (fidelity to the source), and target speakers, who ensure translation fluency. An evaluation showed that collaborative translation improved over raw MT output and back translation, i.e., editing the input to a round-trip machine translation (source-target-source) until the back translation was accepted by the post-editor. Yamashita et al. [49] also considered back translation as a medium for web-based, cross-cultural chat, but did not provide an evaluation.

Hu et al. [23] evaluated iterative refinement of a seed machine translation by pairs of monolinguals. Collaborative translations were consistently rated higher than the original MT output. Hu et al. [24, 25] gave results for other language pairs, with similar improvements in quality. Informal results for time showed that days were required to post-edit fewer than 100 sentences.

MT seed translations might not exist for low-resource language pairs, so Ambati et al. [3] employed weak bilinguals as a bridge between bilingual word translations and monolingual post-editing. Translators with (self-reported) weak ability in either the source or target language provided partial sentence translations, which were then post-edited by monolingual speakers. This staged technique resulted in higher quality translations (according to BLEU [43], an automatic MT metric) on Amazon Mechanical Turk relative to direct solicitation of full sentence translations.



Text from a document about: **a mathematical theory**.

The chance of their doing so is decidedly more

favourable than the chance of the molecules

returning to one half of the vessel.

La chance de leur faire est si nettement plus favorable que la chance des molécules qui reviennent à la moitié du navire.

Feel free to delete the suggested translation if you think that it would be quicker to type your own translation.

Idle: 00:36 (maximum: 03:00)

Figure 2: Web interface for the bilingual post-editing experiment (post-edit condition). We placed the suggested translation in the textbox to minimize scrolling. The idle timer appears on the bottom left.

Experimental Desiderata from Prior Work

Prior published work offers a mixed view[4] on the effectiveness of post-editing due to conflicting experimental designs and objectives. Our experiment clarifies this picture via several design decisions. First, we employ expert bilingual translators, who are faster and more accurate than monolinguals or students. Second, we replicate a standard working environment, avoiding the interference of TAP, eye trackers, and collaborative iterations. Third, we weight time and quality equally, and evaluate quality with a standard ranking technique. Fourth, we assess significance with mixed effects models, which allow us to treat all sampled items (subjects, sentences, and target languages) as random effects. We thus isolate the fixed effect of translation condition, which is the focus of this paper. Finally, we test for other explanatory covariates such as linguistic (e.g., syntactic complexity) and human factors (e.g., source spelling proficiency) features.

EXPERIMENTAL DESIGN

We conducted a language translation experiment with a 2 (translation conditions) x 27 (source sentences) mixed design. Translation conditions (unaided and post-edit), implemented as different user interfaces, and source English sentences were the independent variables (factors). Experimental subjects saw all factor levels, but not all combinations, since one exposure to a source sentence would certainly influence another. We used simple web interfaces (Figure 2) designed to prevent scrolling since subjects worked remotely on their own computers. Source sentences were presented in document order, but subjects could not view the full document context. After submission of each translation, no further revision was allowed. In the post-edit condition, subjects were free to submit, manipulate, or even delete the suggested translation from Google Translate (March 2012). We asked the subjects to eschew alternate machine assistance, although we permitted passive aids like bilingual dictionaries.

[3] In NLP, Callison-Burch [9] investigated monolingual post-editing, but his ultimate objective was improving MT. Both Albrecht et al. [1] and Koehn [30] found that monolingual post-editors could improve the quality of MT output, but that they could not match bilingual translators. Moreover, both found that monolingual post-editors were typically slower than bilingual translators.

[4] Recent, unpublished anecdotal evidence and proprietary trials have more consistently shown the effectiveness of post-editing, motivating adoption at some companies (Ray Flournoy, Adobe, *(p.c.)*).

Subjects completed the experiment under time pressure. Time pressure isolates translation performance from reading comprehension [12] while eliciting a physiological reaction that may increase cognitive function [6]. However, a fixed deadline does not account for per-subject and per-sentence variation, and places an artificial upper bound on translation time. To solve these problems, we displayed an idle timer that prohibited pauses longer than three minutes. The idle timer reset upon any keystroke in the target textbox. Upon expiration, it triggered submission of any entered text. The duration was chosen to allow reflection but to ensure completion of the experiment during a single session.

We recorded all keyboard, mouse, and browser events along with timestamps.[5] The source tokens were also placed in separate `<span>` elements so that we could record hover events.

We randomized the assignment of sentences to translation conditions and the order in which the translation conditions appeared to subjects. Subjects completed a block of sentences in one translation condition, took an untimed break, and then completed the remaining sentences in the other translation condition. Finally, we asked users to complete a questionnaire about the experience.

Selection of Linguistic Materials

We chose English as the source language and Arabic, French, and German as the target languages. The target languages were selected based on canonical word order. Arabic is Verb-Subject-Object (VSO), French is SVO, and German is SOV. Verbs are salient linguistic elements that participate in many syntactic relations, so we wanted to control the position of this variable for cross-linguistic modeling.

We selected four paragraphs from four English Wikipedia articles.[6] We deemed two of the paragraphs "easy" in terms of lexical and syntactic features, and the other two "hard." Subjects saw one easy and one hard document in each translation condition. We selected passages from English articles that had well-written corresponding articles in all target languages. Consequently, subjects could presumably generate natural translations irrespective of the target. Conversely, consider a passage about dating trends in America. This may be difficult to translate into Arabic since dating is not customary in the Arab world. For example, the terms "girlfriend" and "boyfriend" do not have direct translations into Arabic.

The four topics we selected were the 1896 Olympics (easy; Example (1a)), the flag of Japan (easy), Schizophrenia (hard), and the infinite monkey theorem (hard; Example (1b)):

(1) a. It was the first international Olympic Games held in the Modern era.
b. Any physical process that is even less likely than such monkeys' success is effectively impossible, and it may safely be said that such a process will never happen.

	Arabic		**French**		**German**	
	M	SD	M	SD	M	SD
Hourly Rate* ($)	10.34	4.88	17.73	4.37	20.20	10.95
Hours per Week*	31.00	26.13	17.19	13.43	18.88	7.72
En level*	4.94	0.25	4.94	0.25	5.00	0.00
En Skills	4.21	0.34	4.28	0.36	4.34	0.34
En Spelling	4.60	0.42	4.79	0.28	4.78	0.21
En Vocabulary	4.41	0.35	4.40	0.34	4.38	0.55
En-Ar Translation	4.93	0.15				
Fr Spelling			4.72	0.15		
Fr Usage			4.49	0.23		
Fr Vocabulary			4.62	0.22		
En-Fr Translation			4.69	0.19		
De Spelling					4.64	0.30
De Vocabulary					4.68	0.22
En-De Translation					4.77	0.16

Table 1: oDesk human subjects data for Arabic (Ar), English (En), French (Fr), and German (De). oDesk does not currently offer a symmetric inventory of language tests. (*self-reported)

Selection of Subjects

For each target language, we hired 16 self-described "professional" translators on oDesk.[7] Most were freelancers with at least a bachelor's degree. Three had Ph.Ds. We advertised the job at a fixed price of $0.085 per source token ($52 in total), a common rate for general text. However, we allowed translators to submit bids so that they felt fairly compensated. We did not negotiate, but the bids centered close to our target price (mean±standard deviation): Arabic, ($M = 50.50, SD = 4.20$); French, ($M = 52.32, SD = 1.89$); German, ($M = 49.93, SD = 12.57$).

oDesk offers free skills tests administered by a third party.[8] Each 40-minute test contains 40 multiple choice questions, with scores reported on a [0,5] real-valued scale. We required subjects to complete all available source and target language proficiency tests, in addition to language-pair-specific translation tests. We also recorded public profile information such as hourly rate, hours worked as a translator, and self-reported English proficiency. Table 1 summarizes the subject data.

Subjects completed a training module that explained the experimental procedure and exposed them to both translation conditions. They could translate example sentences until they were ready to start the experiment.

Translation Quality Assessment

Translation quality assessment is far from a solved problem. Whereas past experiments in the HCI community have used 5-point fluency/adequacy scales, the MT community has lately favored pairwise ranking [11]. Pairwise ranking results in higher inner annotator agreement (IAA) than fluency/adequacy rating [11]. We used software[9] from the annual Workshop on Machine Translation (WMT) evaluations to rank all translations on Amazon Mechanical Turk (Figure 3). Aggregate non-expert judgements can approach expert IAA levels [10].

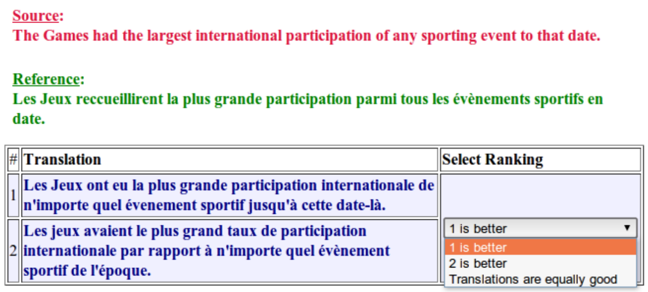
[5] We did not record cut/copy/paste events [13]. Analysis showed that these events would be useful to track in future experiments.

[6] Gross statistics: 27 sentences, 606 tokens. The maximum sentence length was 43, and the average length was 22.4 tokens.

[7] `http://www.odesk.com`

[8] ExpertRating: `http://www.expertrating.com`

[9] `http://cs.jhu.edu/~ozaidan/maise/`



Source:
The Games had the largest international participation of any sporting event to that date.

Reference:
Les Jeux reccueillirent la plus grande participation parmi tous les évènements sportifs en date.

#	Translation	Select Ranking
1	Les Jeux ont eu la plus grande participation internationale de n'importe quel évenement sportif jusqu'à cette date-là.	
2	Les jeux avaient le plus grand taux de participation internationale par rapport à n'importe quel évènement sportif de l'époque.	1 is better ▾ 1 is better 2 is better Translations are equally good

Figure 3: Three-way ranking interface for assessing translation quality using Amazon Mechanical Turk. Raters could see the source sentence, a human-generated reference translation, and the two target sentences. Each HIT contained three ranking tasks.

The combination of pairwise rankings into a total ordering is non-trivial. Currently, WMT casts the problem as finding the minimum feedback arc set in a tournament (a directed graph with $\binom{N}{2}$ vertices, where N is the number of human translators). This is an NP-complete problem, but it can be approximated with the algorithm of Lopez [36].[10]

We performed an exhaustive pairwise evaluation of the translation results for all three languages. For each source sentence, we requested three independent judgements for each of the $\binom{N}{2}$ translation pairs. We paid \$0.04 per HIT, and each HIT contained three pairs. Workers needed an 85% approval rating with at least five approved HITs. For quality control, we randomly interspersed non-sensical HITs—the translations did not correspond to the source text—among the real HITs. We blocked workers who incorrectly answered several spam HITs.

Raters were asked to choose the best translation, or to mark the two translations as equal (Figure 3). For each source sentence and each set of N target translations, the procedure resulted in a ranking in the range [1,N], with ties allowed.

VISUALIZING TRANSLATION ACTIVITY

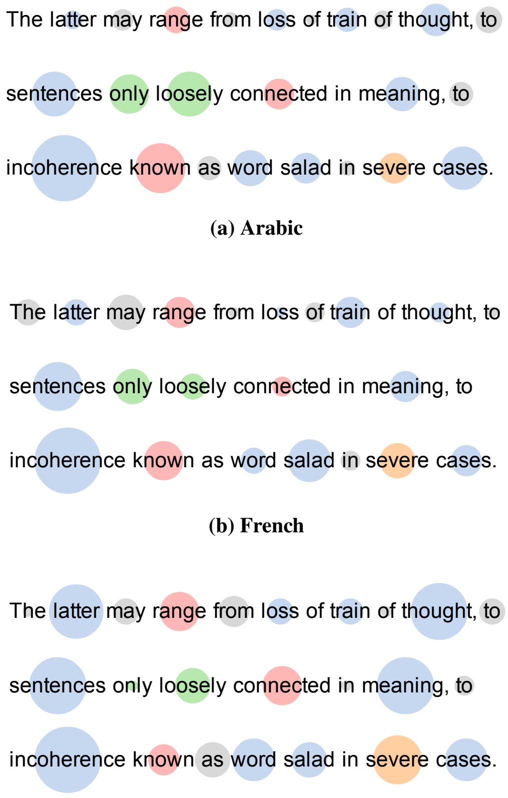
We visualized the user activity data to assess observations from prior work and to inform the statistical models.

Mouse Cursor Movements

Figure 4 shows an example English sentence from the Schizophrenia document with hover counts from all three languages. The areas of the circles are proportional to the square root of the raw counts, while the colors indicate the various parts of speech: noun, verb, adjective, adverb, and “other”. The “other” category includes prepositions, particles, conjunctions, and other (mostly closed) word classes.

Nouns stand out as a significant focal point, as do adverbs and, to a lesser degree, verbs. These patterns persist across all three languages, and suggest that source parts of speech might have an effect on time and quality. We assess that hypothesis statistically in the next section.

Huang et al. [26] showed that mouse movements correlated with gaze for search engine UIs. While we did not correlate mouse movements with an eye tracker for our task, the visualization nonetheless shows distinctive patterns that turn out to be significant in our statistical models.



(a) Arabic

(b) French

(c) German

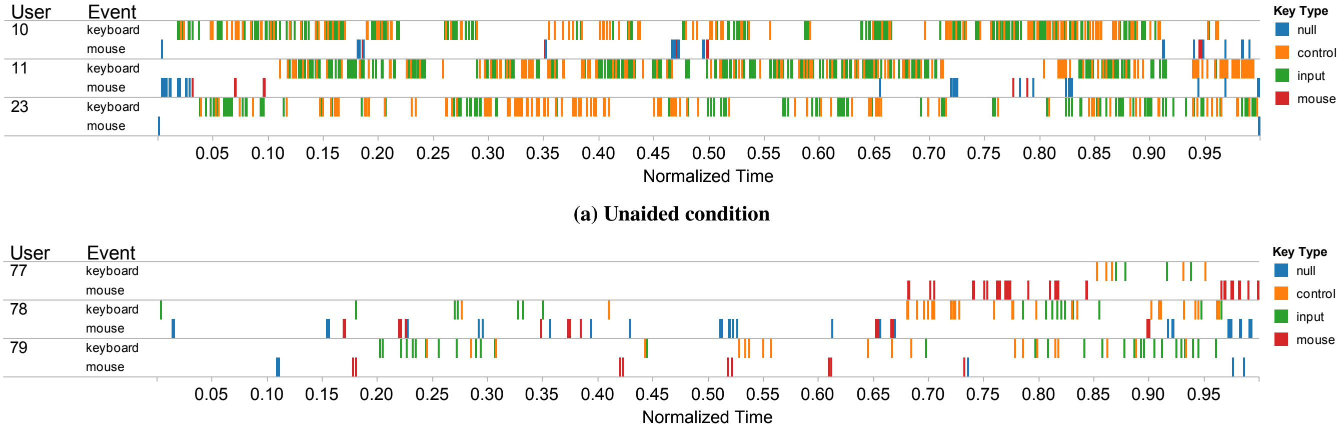
Figure 4: Mouse hover frequencies for the three different languages. Frequency is indicated by area, while the colors indicate five word categories: nouns (blue), verbs (red), adjectives (orange), adverbs (green), and “other” (grey). Nouns are clearly focal points.

User Event Traces

We also plotted the mouse and keyboard event logs against a normalized time scale for each user and source sentence (Figure 5). Users in the unaided condition demonstrate the gisting/drafting/revising behavior observed in prior work with eye trackers [15]. Initial pauses and mouse activity in the gisting phase give way to concentrated keyboard activity as the user types the translation. Finally, more pauses and mouse activity indicate the revision phase.

The post-edit condition results in drastically different behavior. Phase boundaries are not discernible, and pauses account for a larger proportion of the translation time. Users clearly engaged the suggested translation even though the option to discard it existed. In addition, the post-edit condition resulted in a statistically significant reduction in total event counts: Arabic $t(26) = 16.52, p < 0.001$; French $t(26) = 33.63, p \leq 0.001$; German $t(26) = 37.08, p < 0.001$. At least from the perspective of device input activity, post-editing is a more passive activity than unaided translation.

[10] `http://github.com/alopez/wmt-ranking`



(a) Unaided condition

(b) Post-edit condition

Figure 5: Arabic user activity logs for the English input shown in Figure 4. Events are colored according to input type: control keys (e.g., up/down arrows; orange), input keys (green), mouse buttons (red), and "null" (blue), which means that Javascript failed to trap the input character. (a) The unaided condition results in visible gisting, drafting, and revision phases, which are sequential. (b) The post-edit condition does not result in clear phases. Pauses are longer and account for more of the total translation time.

STATISTICAL ANALYSIS

A consistent observation in prior work on MT post-editing has been significant per-subject and per-source-sentence (in terms of the quality of the machine suggestion) variation. Two major statistical mistakes have been made in dealing with this variance. First, Tatsumi [47] used a multiple regression design with predictors for each subject. This is the least effective solution for regression—per-subject regressions or by-item means would be superior [4]—since her model measured significance of each subject relative to the others, but not to the larger human population.

An alternative to multiple regression is analysis of variance (ANOVA), the workhorse of HCI experimental layouts. Per-subject variance could be modeled with a repeated measures design (RM-ANOVA), but our experiment was necessarily a between-subjects design. Nonetheless, we could treat source sentences as a fixed effect and use a two-way, between-subjects ANOVA. However, language is not a fixed effect since we cannot include all possible source sentences in the experiment. The "language-as-fixed-effects fallacy" [18] has been the motivation in psycho-linguistics research for *mixed effects models*, which can incorporate arbitrary fixed effects (e.g., translation conditions) and random effects (e.g., subjects, sentences).[11]

Mixed effects models are not widely used in HCI, but they are preferable to ANOVA whenever experimental factors are sampled from a larger population. In this setting, mixed models reduce the risk of type II errors. Other practical benefits include accommodation of unbalanced designs and longitudinal (dependent) covariates, which can model learning and fatigue. We assess significance with likelihood-ratio (LR) tests.

Translation Time and Quality

This paper seeks to quantify the effect of post-editing on translation time and quality (Figure 6).[12] Plots showed that time was not normally distributed, so we applied a log transformation and fit linear mixed models.[13] As for quality, our ranking procedure yields discrete ranks, which are obviously not normally distributed. To account for non-normality and to exploit the ordering of the rank response, we built ordinal mixed models.[14]

In addition to the per-subject covariates listed in Table 1, we also included source (English) features. We annotated the source text with Stanford CoreNLP[15] to obtain syntactic complexity [34], number of named entities, and part of speech tag counts. We also included (log transformed) mean lexical frequency[16] and sentence length. We standardized all numeric covariates by subtracting the mean and dividing by one standard deviation [21]. Finally, after fitting models with all of the data, we pruned data points with residuals exceeding 2.5 standard deviations [4] and refit the models.

Monolingual Time and Quality Results

Table 2 shows results for log-transformed time. The standard two-way ANOVA model, which treats source sentences as fixed effects, shows a significant main effect for translation condition across all three languages. However, the risk of type II errors is high, since the ANOVA does not account for per-subject or per-sentence variance. The mixed effects model addresses

[11] See Appendix for mathematical details.

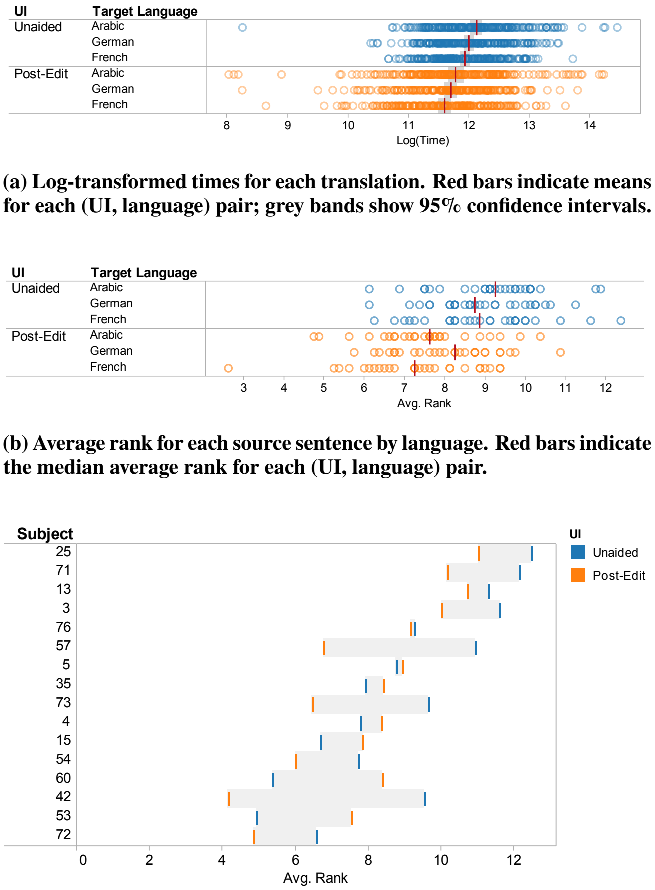
[12] Intuition suggests that time and quality might be correlated. For each language, we ran Pearson's correlation test for the two response variables: Arabic, $t(430) = 0.739, p < 0.274$; French, $t(430) = 3.71, p < 0.001$; German, $t(430) = 2.49, p < 0.05$. Positive correlations—monotonic increases in time correspond to increases in rank (lower quality)—exist for two languages. Consequently, multivariate models might be a direction for future research.

[13] With the `lme4` R package [5].

[14] With the `ordinal` R package [17].

[15] **`http://nlp.stanford.edu/software/corenlp.shtml`**

[16] Relative to counts in the English Gigaword corpus (Linguistic Data Consortium catalog number LDC2009T13).



(a) Log-transformed times for each translation. Red bars indicate means for each (UI, language) pair; grey bands show 95% confidence intervals.

(b) Average rank for each source sentence by language. Red bars indicate the median average rank for each (UI, language) pair.

(c) Average rank in each condition by subject (German).

Figure 6: Summarizations of the post-editing experiment data. (a) Mean time is lower for post-edit. (b) Median rank is better for post-edit. (c) Individual responses show high variance. German subjects 4, 5, 15, 35, 53, and 60 are actually faster, on average, in the unaided condition.

Model	Factor	Ar	Fr	De	ALL
ANOVA	**post-edit**	•••	•••	•••	–
	source sentence	•••	•••	•••	–
Mixed Effects	**post-edit**	••	••	•••	•••
	Hourly Rate* (\$)	•			
	En level*		•		
	log source length	•••	•••	•••	•••
	log syntax complexity	•			
	% nouns in source	••	•	•	••
	Translation Test	•			–
	Target Spelling Test			•	–

Table 2: Time results for post-edit and other significant covariates. Insignificant covariates in the full model are not listed. Statistical significance: ••• $p < 0.001$; •• $p < 0.01$; • $p < 0.05$.

Model	Factor	Ar	Fr	De	ALL
Mixed Effects	**post-edit**	•••	•••	•	•••
	Hourly Rate* (\$)	•			
	En Skills Test		•		•
	Target Vocab test		•••		–

Table 3: Rank results for post-edit and other significant covariates.

these shortcomings with random intercepts for subjects and sentences, and random slopes for translation condition. Informally, the random slopes can be interpreted as follows: translation condition might affect some subjects more than others, and its affect might also depend on the source sentence (Figure 6c). These interpretations account for observations in prior work, namely that MT suggestions help some subjects more than others, and that MT quality varies across source sentences.

The mixed models found significant main effects for translation condition in Arabic $\chi^2(1, N = 432) = 7.54, p < 0.01$, French $\chi^2(1, N = 432) = 10.12, p < 0.01$, and German $\chi^2(1, N = 432) = 19.69, p < 0.001$. [17]

For rank (Table 3), we included random intercepts for subjects and sentences (`ordinal` [17] does not yet support random slopes) in the ordinal models. We found significant main effects for translation condition in Arabic $\chi^2(1, N = 432) = 16.09, p < 0.001$, French $\chi^2(1, N = 432) = 15.24, p < 0.001$, and German $\chi^2(1, N = 432) = 4.03, p < 0.05$.

Cross-lingual Time and Quality Results

We also pooled the data for all three languages and built new mixed effects models with target language as an additional random effect. Effects not present in the monolingual models may be detected in the cross-lingual model due to increased statistical power. A three-way ANOVA with target language as a fixed effect is possible yet clearly flawed, so we omit it. We also removed the translation and target language test covariates since subjects only took the tests for their target language. For time, we found a significant main effect for translation condition $\chi^2(1, N = 1296) = 11.58, p < 0.001$ ("ALL" in Tables 2 and 3). Source sentence length and the percentage of source noun tokens were significant covariates. For rank, we also found a significant main effect for translation condition $\chi^2(1, N = 1296) = 32.47, p < 0.001$.

Translation Edit Distance

We have shown that translation condition has a significant effect on quality. Does the suggested translation prime the post-editor, thus resulting in a translation that is similar to the suggestion? We computed the Damerau-Levenshtein edit distance[18] between the suggested translation and each target translation, and then averaged the edit distances for each source sentence and translation condition. We then used a paired difference t-test to assess significance. We found statistically significant reductions in edit distance for the post-edit condition: Arabic $t(26) = 68.45, p < 0.001$; French $t(26) = 39.02, p < 0.001$; German $t(26) = 55.65, p < 0.001$.

[17] The data set and scripts to reproduce these experiments (with effect sizes) are available at: **`http://www.spencegreen.com/research/`**

[18] Extends Levenshtein edit distance with transpositions to account for word order differences.

Response	Likelihood-ratio Test		Sign
Count (300ms)	$\chi^2(1, N = 1296) = 10.22$	$p < 0.01$	$-$
Count (1000ms)	$\chi^2(1, N = 1296) = 7.12$	$p < 0.01$	$-$
Mean duration	$\chi^2(1, N = 1296) = 9.37$	$p < 0.01$	$+$
Ratio (300ms)	$\chi^2(1, N = 1296) = 4.87$	$p < 0.05$	$+$
Ratio (1000ms)	$\chi^2(1, N = 1296) = 11.06$	$p < 0.001$	$+$

Table 4: Effect of translation condition on pause count, duration, and ratio (all languages). "Sign" refers to the polarity of the fixed co-efficient for post-edit. The sign indicates that post-editing results in fewer distinct pauses (count), that are longer (mean duration), and account for more of the total translation time (ratio).

Pause Duration and Frequency

In translation process studies, pauses have frequently been used as proxies for cognitive load. Schilperoord [45] used a cutoff of 300ms between input events, while others [29, 42] have used 1000ms. We fit the same mixed effects models used for time to the following response variables: pause count (300ms and 1000ms), mean pause duration (300ms), and pause ratio (300ms and 1000ms). Pause ratio is the total pause time divided by translation time.

We logged all events in the DOM Level 3 specification except: `abort`, `keyup`, `mousemove`, and `wheel`. Pauses are the intervals between these events, a slightly broader definition than past work (see [29, 42]).

Due to space constraints, we only report cross-lingual significance levels for translation condition. Table 4 shows that a significant main effect for the post-edit condition exists for all five response variables. The polarity of the co-efficients indicates that post-edit results in fewer total pauses. These pauses are longer and account for a larger fraction of the total translation time. These results support the differences in user behavior observed in the event visualizations (Figure 5).

DISCUSSION

The triangulation of user activity visualization, statistical modeling, and qualitative user feedback yields new insights into machine-assisted translation.

Translators strongly prefer post-edit

Translators have previously reported a strong dislike of full MT pre-translations [32, 48]. However, our cross-linguistic mixed effects model showed significant main effects for both time and quality. Visualization of the user activity logs showed fewer interactions and longer pauses in the post-edit condition, observations that we verified with mixed effects models (Table 4). Moreover, when asked, "Which interface enabled you to translate the sentences most quickly?", 69% of translators chose post-editing. When asked, "Were the machine translations useful?", 56% responded in the affirmative, 29% were neutral, and only 15% disagreed. One user even responded,

```
Your machine translations are far better than
the ones of Google, Babel and so on.  So they
wered helpfull [sic], but usually when handed
over google-translated material, I find it way
easier end [sic] quicker to do it on my own
from unaided.
```

The subjects did not know that the suggestions came from Google Translate. Users may have dated perceptions of MT quality that do not account for the rapid progress in the field.

Post-edit has measurably different interaction patterns

Post-editing seems like a more passive activity from several angles. We found statistically significant reductions in event counts, visual (Figure 5) and statistical (Table 4) evidence of longer duration pauses, and final translations that were similar to the suggestions. Users did not devote as much time to text generation (Figure 5b). However, we did not find general human factors predictors of translation time or quality. For example, only a few of the oDesk predictors in Table 1 were significant in the monolingual mixed models, and none of them were significant in the cross-lingual models. Of course, we recruited experts, so the scores were near perfect with low variance. Nonetheless, post-edit still had a significant per subject random component, as illustrated by the individual variation in Figure 6c. We used random slopes to account for this variance, but design principles are difficult to infer from random effects. Closer human subjects analysis is needed to isolate fixed predictors of post-edit productivity.

Suggested translations improve final quality

Across languages, we found a very significant main effect ($p < 0.001$) for quality. Machine translation quality has advanced since many earlier post-editing studies were conducted. Nonetheless, even recent research offers a mixed view on the products of post-editing, with [22, 46] reporting lower quality and [20, 44] reporting higher quality. We found very significant effects ($p < 0.001$) for Arabic and French, and a less significant effect ($p < 0.05$) for German. The user survey suggests that the German translators may have been optimizing for time instead of quality. When asked "Which interface enabled you to translate the sentences most quickly?", 75% chose the post-edit condition, the highest proportion among the three language groups. When asked if the suggested translations were "useful," 50% answered in the affirmative, 37% were neutral, and only 12.5% disagreed.

Simple source lexical features predict time

Prior work measured the effect of rule-based, language-specific source-side features [7, 39, 40] that may not generalize across languages and text genres. In contrast, we found that very general linguistic features such as basic part of speech counts and syntactic complexity predict translation time. Across languages, we found a significant main effect for %nouns, the proportion of nominal tokens in the source sentence. One notable omission from this list is %verbs, which we hypothesized would be a salient linguistic feature. The mouse hover patterns showed that users fixated on nouns, and to a lesser degree adjectives. However, the user survey presented a slightly different picture. Across languages, users provided the following ranking of basic parts of speech in order of decreasing translation difficulty: Adverb, Verb, Adjective, Other, Noun.[19] Translators seem aware of the difficulty of adverbs, but apparently underestimate the difficulty of nouns.

[19] The users who ranked "Other" highly were asked to further qualify their response. These responses uniformly demonstrate a basic misunderstanding of the concept of part of speech.

UI DESIGN IMPLICATIONS

Our results and analysis suggest several design principles for new and existing machine-assisted translation systems.

Show Translations for Selected Parts of Speech

Both the activity visualizations and mixed effects models indicate that users struggle with certain parts of speech, such as nouns and adverbs. Verbs, prepositions, and other parts of speech did not affect translation time. While many translation interfaces support dictionary lookup, users may benefit from the automatic display of translations for certain parts of speech.

Avoid Predicting Translation Modes

Conventional wisdom has it that translation consists of gisting, drafting, and revising phases [15]. However, the user activity logs show that these phases are interleaved in the post-edit condition. The system should not be tuned to a specific mode. One option would be to allow the user to toggle specific assistance for each translation phase.

Offer Full Translations as References

We used a simple interface in which the suggested translation appeared in the input textbox. Several translators commented that they pasted the translation elsewhere for reference, using only parts of it in their final submission. The activity visualizations also indicated a significant amount of control key activity in the post-edit condition, meaning that users were navigating with arrow keys through the suggestion. We conclude that the suggested translation should be accessible, but should not impede entry of the final translation.

Use Post-Edit Translations to Improve MT

This paper has concerned itself with assistance to the human translator. However, the edit distance analysis showed that translations produced in the post-edit condition diverged from both the unaided translations and the raw MT output, yet were closer to MT: humans start from MT, then produce a novel output. The words, phrases, and re-orderings applied in that production process could be used as additional training data, thus creating a virtuous cycle in which both humans and machines improve. UI design can play a role in this loop by automating manipulation of the machine assistance, thus encouraging user actions that can be recorded. For example, we have mentioned the utility of the full machine translation as a reference. Instead of allowing uncontrolled copy/paste operations, a UI feature could automate selection of words and phrases. An event trace of these edit operations could be used as features in existing discriminative MT training algorithms. We suggest that UI designers and MT system builders work cooperatively to improve both human and machine translation.

CONCLUSION

Natural language processing systems have entered mainstream use, yet few analyses of human interaction with them exist. Machine translation systems are among the most widely used, and they hold great promise for lowering the high cost of translation. We analyzed MT post-editing, a common feature in commercial translator interfaces. Our results strongly favor the presence of machine suggestions in terms of both translation time and final quality. If translators benefit from a barebones post-editing interface, then we suspect that more interaction between the UI and MT backend could produce additional benefits. While Koehn [29] found that interactive assistance offered no advantage over post-editing, he was less concerned about UI design and human factors. Our results fill that gap, and inform the design of new interactive interfaces, which should be compared to post-editing. Our work also applies directly to existing translator workbenches.

APPENDIX: FROM ANOVA TO MIXED EFFECTS MODELS

Consider a balanced, one-way analysis of variance (ANOVA) model with $i \in \{1, \dots, I\}$ clusters (e.g., translation conditions) and $j \in \{1, \dots, J\}$ observations per cluster:

$$y_{ij} = x_{ij}\beta_i + \epsilon_{ij} \tag{1}$$

where y are normally distributed observations (dependent variables), x are the independent variables (effects), the β_i are the group means for $y_{i\cdot}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ are random per-observation errors (residuals) with common variance. The null hypothesis is that the co-efficients, which are fit with ordinary least squares, are equal for all clusters, i.e. $\beta_1 = \beta_2 = \cdots = \beta_I$. The F-test for this model is a ratio of the between- and within-subjects sum-of-squared (SS) errors, which has the familiar form

$$F = \frac{SS_b(IJ - I)}{SS_w(IJ - 1)} \tag{2}$$

Random differences among clusters are not modeled explicitly, thus increasing the risk of type II errors. One method for separating the per-cluster/subject error is a repeated measures design (RM-ANOVA),[20] which is widely used in HCI. RM-ANOVA differs only in terms of SS_w, which is partitioned into subject-specific error plus some residual error:

$$SS_w = SS_{subjects} + SS_{error} \tag{3}$$

RM-ANOVA effectively adds a per-subject random effect without changing the form of the model in Equation 1. The F-test becomes:

$$F = \frac{SS_b(I-1)(IJ-1)}{SS_{error}(IJ-1)} = \frac{SS_b(I-1)}{SS_{error}} \tag{4}$$

The F-value of Equation 4 will increase if SS_{error} is sufficiently small relative to SS_w (Equation 2) so as to offset the loss in degrees of freedom, thus increasing statistical power.

What if we want to include other random effects like sentences or target languages? *Mixed models* extend the ANOVA model in Equation 1 with an explicit random effects component:

$$y_i = x_i^{\intercal}\beta + z_i^{\intercal} b_i + \eta_i \tag{5}$$

where i is now called the *grouping factor*. We now write y_i as a $J \times 1$ vector of responses for the ith cluster/group and x_i as a $J \times p$ vector of covariates with fixed co-efficients $\beta \in \mathbb{R}^{p \times 1}$. Further, $E[y_i] = x_i^{\intercal}\beta$. The random effects structure is defined by a $J \times q$ vector of covariates z_i with associated random effects $b_i \sim \mathcal{N}(0, \Sigma)$, $\Sigma \in \mathbb{R}^{q \times q}$ and the residuals $\eta_i \sim \mathcal{N}(0, R_t)$, $R_t \in \mathbb{R}^{I \times I}$. The model assumes mutual independence of all random vectors.

The vectors x_i and z_i relate the observations y_i to β and b_i. For significance, F-tests no longer apply, but we can use a likelihood-ratio (LR) test for specific fixed effects. Our analysis used LR testing, so we fit the mixed models by maximizing log-likelihood.

ACKNOWLEDGEMENTS

We thank Ray Flournoy and Roger Levy for helpful discussions. The first author is supported by a National Science Foundation Graduate Research Fellowship.

[20] Also known as a "within-subjects" experimental design.

REFERENCES

1. Albrecht, J. S., Hwa, R., and Marai, G. E. Correcting automatic translations through collaborations between MT and monolingual target-language users. In *EACL* (2009).

2. Allen, J. Post-editing. In *Computers and Translation: A translator's guide*, H. Somers, Ed. John Benjamins, 2003, ch. 16, 297–317.

3. Ambati, V., Vogel, S., and Carbonell, J. Collaborative workflow for crowdsourcing translation. In *CSCW* (2012).

4. Baayen, R., Davidson, D., and Bates, D. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language 59*, 4 (2008), 390–412.

5. Bates, D. M. `lme4`: Linear mixed-effects models using S4 classes. Tech. rep., R package version 0.999999-0, **`http://cran.r-project.org/package=lme4`**, 2007.

6. Bayer-Hohenwarter, G. Methodological reflections on the experimental design of time-pressure studies. *Across Languages and Cultures 10*, 2 (2009), 193–206.

7. Bernth, A., and McCord, M. The effect of source analysis on translation confidence. In *Envisioning Machine Translation in the Information Future*, J. White, Ed., vol. 1934 of *Lecture Notes in Computer Science*. Springer, 2000, 250–259.

8. Bisbey, R., and Kay. The MIND translation system: a study in man-machine collaboration. Tech. Rep. P-4786, Rand Corp., March 1972.

9. Callison-Burch, C. Linear B system description for the 2005 NIST MT evaluation exercise. In *NIST Machine Translation Evaluation Workshop* (2005).

10. Callison-Burch, C. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *EMNLP* (2009).

11. Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. Findings of the 2012 Workshop on Statistical Machine Translation. In *WMT* (2012).

12. Campbell, S. A cognitive approach to source text difficulty in translation. *Target 11*, 1 (1999), 33–63.

13. Carl, M. Translog-II: a program for recording user activity data for empirical reading and writing research. In *LREC* (2012).

14. Carl, M., and Jakobsen, A. Towards statistical modelling of translators' activity data. *International Journal of Speech Technology 12* (2009), 125–138.

15. Carl, M., Kay, M., and Jensen, K. T. H. Long distance revisions in drafting and post-editing. In *CICLing* (2010).

16. Chen, M. C., Anderson, J. R., and Sohn, M. H. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI* (2001).

17. Christensen, R. H. B. `ordinal`: Regression models for ordinal data. Tech. rep., R package version 2012.01-19, **`http://cran.r-project.org/package=ordinal`**, 2011.

18. Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior 12*, 4 (August 1973), 335–359.

19. Flournoy, R., and Duran, C. Machine translation and document localization at Adobe: From pilot to production. In *MT Summit XII* (2009).

20. Garcia, I. Translating by post-editing: is it the way forward? *Machine Translation 25* (2011), 217–237.

21. Gelman, A. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine 27*, 15 (2008), 2865–2873.

22. Guerberof, A. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *International Journal of Localization 7*, 1 (2009).

23. Hu, C., Bederson, B., and Resnik, P. Translation by iterative collaboration between monolingual users. In *Graphics Interface (GI)* (2010).

24. Hu, C., Bederson, B. B., Resnik, P., and Kronrod, Y. MonoTrans2: a new human computation system to support monolingual translation. In *CHI* (2011).

25. Hu, C., Resnik, P., Kronrod, Y., and Bederson, B. Deploying MonoTrans widgets in the wild. In *CHI* (2012).

26. Huang, J., White, R., and Buscher, G. User see, user point: gaze and cursor alignment in web search. In *CHI* (2012).

27. Jakobsen, A. Logging target text production with Translog. In *Probing the process in translation: methods and results*, G. Hansen, Ed. Copenhagen: Samfundslitteratur, 1999, 9–20.

28. Kay, M. The proper place of men and machines in language translation. Tech. Rep. CSL-80-11, Xerox Palo Alto Research Center (PARC), 1980.

29. Koehn, P. A process study of computer-aided translation. *Machine Translation 23* (2009), 241–263.

30. Koehn, P. Enabling monolingual translators: post-editing vs. options. In *HLT-NAACL* (2010).

31. Krings, H. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent State University Press, 2001.

32. Lagoudaki, E. Translation editing environments. In *MT Summit XII: Workshop on Beyond Translation Memories* (2009).

33. Lawson, V., Ed. *Practical Experience of Machine Translation*, North-Holland (1982).

34. Lin, D. On the structural complexity of natural language sentences. In *COLING* (1996).

35. Lopez, A. Statistical machine translation. *ACM Computing Surveys 40*, 8 (2008), 1–49.

36. Lopez, A. Putting human assessments of machine translation systems in order. In *WMT* (2012).

37. Morita, D., and Ishida, T. Collaborative translation by monolinguals with machine translators. In *IUI* (2009).

38. Morita, D., and Ishida, T. Designing protocols for collaborative translation. In *Principles of Practice in Multi-Agent Systems* (2009).

39. O'Brien, S. Machine translatability and post-editing effort: How do they relate? In *Translating and the Computer* (2004).

40. O'Brien, S. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation 19* (2005), 37–58.

41. O'Brien, S. Eye-tracking and translation memory matches. *Perspectives: Studies in translatology 14*, 3 (2006), 185–205.

42. O'Brien, S. Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures 7*, 1 (2006), 1–21.

43. Papineni, K., Roukos, S., Ward, T., and Zhu, W. BLEU: a method for automatic evaluation of machine translation. In *ACL* (2002).

44. Plitt, M., and Masselot, F. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics 93* (2010), 7–16.

45. Schilperoord, J. *It's About Time: Temporal Aspects of Cognitive Processes in Text Production*. Amsterdam:Rodopi, 1996.

46. Skadiņš, R., Puriņš, M., Skadiņa, I., and Vasiļjevs, A. Evaluation of SMT in localization to under-resourced inflected language. In *EAMT* (2011).

47. Tatsumi, M. *Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis*. PhD thesis, Dublin City University, 2010.

48. Wallis, J. Interactive translation vs pre-translation in the context of translation memory systems: Investigating the effects of translation method on productivity, quality and translator satisfaction. Master's thesis, University of Ottawa, 2006.

49. Yamashita, N., Inaba, R., Kuzuoka, H., and Ishida, T. Difficulties in establishing common ground in multiparty groups using machine translation. In *CHI* (2009).